MATH337: Changepoint Detection

Gaetano Romano

2024-11-01

Table of contents

Preface

These are the notes for **MATH337 Changepoint Detection**. They were written by Gaetano Romano.

The module will introduce you to changepoint detection, detailing some algorithms, developing the basics theoretical foundations, and practicing few real-world scenarios.

Across five weeks we will cover the following topics:

1. An introduction to changepoint detection and the CUSUM statistics
2. Controlling the CUSUM and some additional models
3. Dealing with multiple changes
4. PELT, WBS and Penalty selection
5. Working with Real World data.

We will be using R as the programming language for this module. If you're unfamiliar with it, make sure you cover the first three weeks of MATH245.

Every week, you are expected to follow two lectures, one workshop, and one computer aided lab. Over the lecture, we will cover the basics concepts of changepoint detection.

At the end of each chapter, you will find exercises that will be carried in the workshop and the lab. During the workshop, you will be dealing with computations and details about the methodologies, and, finally, during the lab sessions, you'll give a go at programming the various algorithms and running real-world examples.

You will find the solutions to the exercises on the Moodle page, released weekly. If you cannot access the Moodle page, and you still would like to have these solutions, please get in touch with me.

Source files, and attributions

The materials in this course are based on and share elements with the following resources:

- Fearnhead, P., & Fryzlewicz, P. (2022). Detecting a single change-point. *arXiv preprint arXiv:2210.07066*.
- Rebecca Killick's Introduction to Changepoint Detection - a half-day introductory course on changepoint detection.
- Rebecca Killick's Further Changepoint Topics - an extended course on changepoint detection.
- Toby Hocking's Course on Unsupervised Learning, which includes changepoint detection.

I would like to express my gratitude to the authors of these resources. In addition, materials were sourced from various academic papers, which are referenced throughout the body of these notes.

1 An Introduction to Changepoint Detection

1.1 Piecewise Stationary Time Series

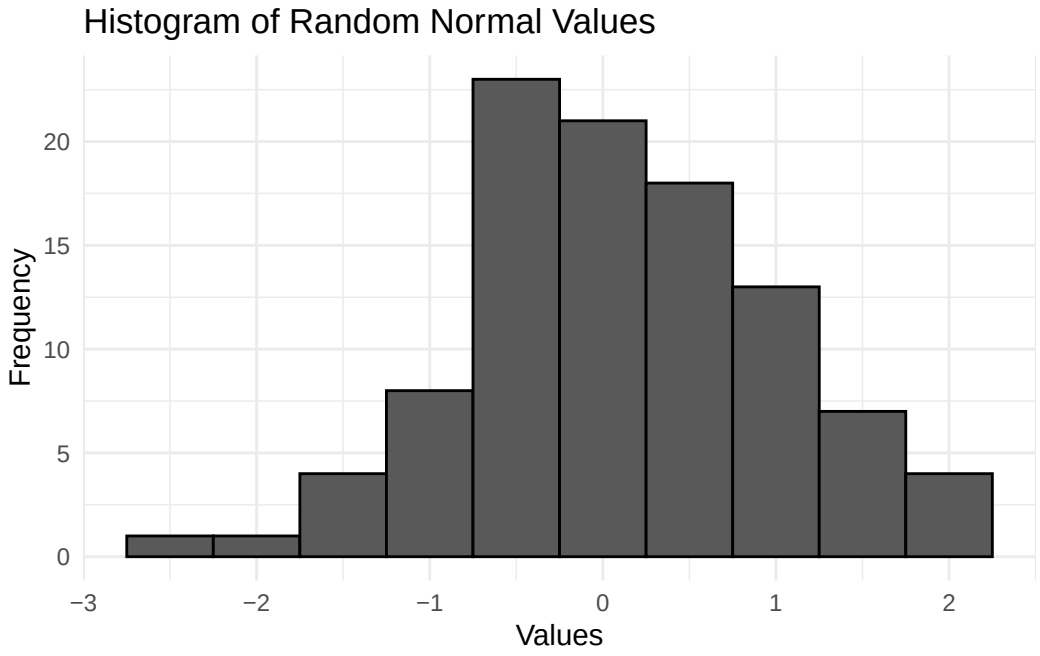
In this module, we will be dealing with **time series**. A time series is a sequence of observations recorded over time (or space), where the order of the data points is crucial.

1.1.1 What is a time series?

In previous modules, such as Likelihood Inference, we typically dealt with data that was not ordered in a particular way. For example, we might have worked with a sample of independent Gaussian observations, where each observation is drawn randomly from the same distribution. This sample might look like the following:

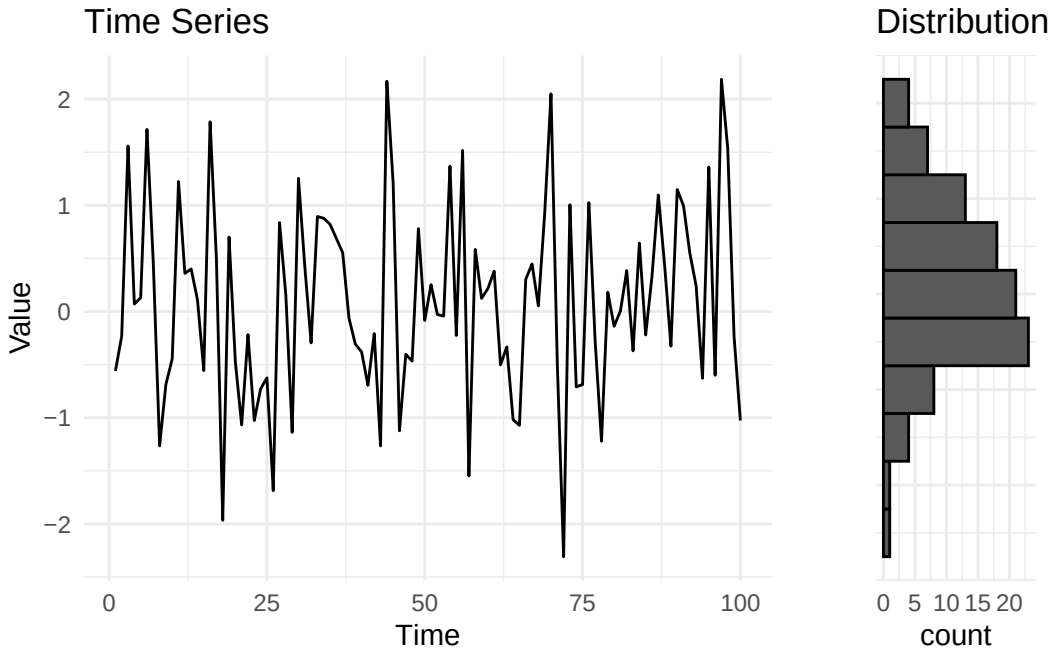
$$y_i \sim \mathcal{N}(0, 1),\ i = 1, \ldots, 100$$

Here, y_i represents the i-th observation, and the assumption is that all observations are independent and identically distributed (i.i.d.) with a mean of 0 and variance of 1.



In this case, the observations do not have any particular order, and our primary interest may be in estimating parameters such as the mean, variance, or mode of the distribution. This is typical for traditional inference, where the order of observations is not of concern.

However, a **time series** involves a specific order to the data—usually indexed by time, although it could also be by space or another sequential dimension. For example, we could assume that the Gaussian sample above is a sequential process, ordered by the time we drew an observation. Each observation corresponds to a specific time point t.



Formal Notation. In time series analysis, use an index t to represent time or order on a given set of observations. The time series vector is written as:

$$y_{1:n} = (y_1, y_2, \dots, y_n).$$

Here, n is the total length of the sequence, and y_t represents the observed value at time t, for $t = 1, 2, \dots, n$. In our previous example, for instance, $n = 100$.

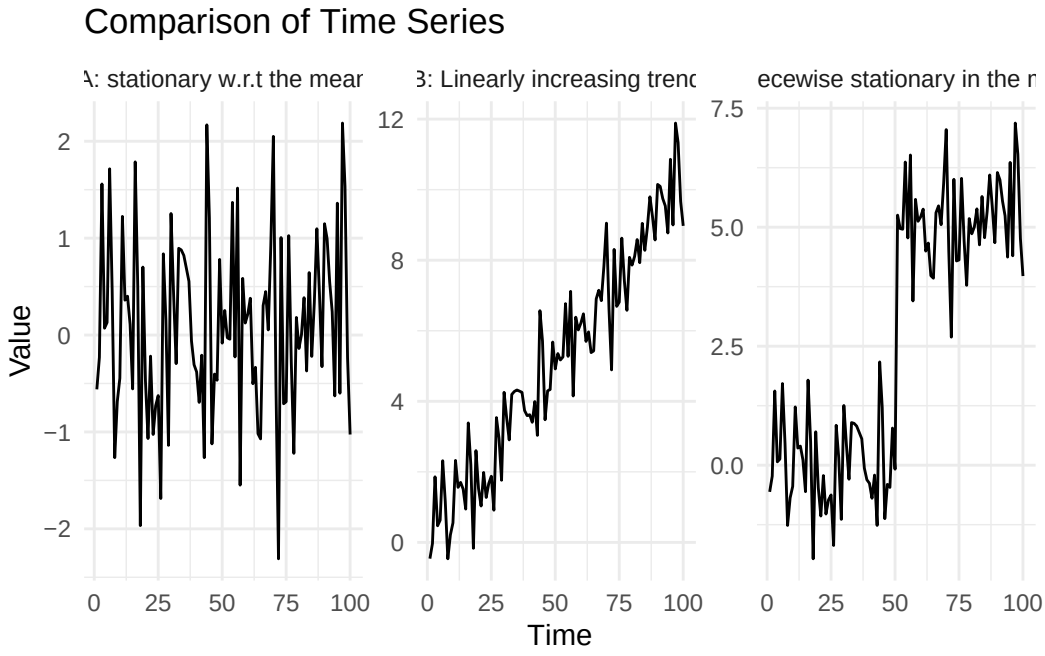
Often, we are also interested in subsets of a time series, especially when investigating specific "windows" or "chunks" of the data. A subset of a time series, starting from time l to time u, will be denoted by the following:

$$y_{l:u} = (y_l, y_{l+1}, \dots, y_u).$$

Understanding and working with subsets of time series data is important for many applications, such as when detecting changes in the behavior or properties of the time series over specific intervals.

1.1.2 Stationary, non-stationary, and piecewise stationary time series

Time series can have various statistical properties that explain how they behave over time, and they can be characterized based on those. Let us look at three examples of time series:



A. The leftmost time series, was generated by sampling random normal variables $y_t = \epsilon_t,\ \epsilon_t \sim \mathcal{N}(0,1)$. In this case:

$$\mathbb{E}(y_t) = \mathbb{E}(\epsilon_t) = 0,\ \mathrm{Var}(y_t) = \mathrm{Var}(\epsilon_t) = 1,\ \forall\ t \in \{1, ..., 100\}.$$

Say we generate more observations under the same random process, this will give us still a value that will be centered on 0, with variance 1, e.g. $\mathbb{E}(y_{150}) = 0,\ \mathrm{Var}(y_{150}) = 1$.

B. In the centre time series, the series is generated as:

$$y_t = \epsilon_t + 0.1 \cdot t,\ \epsilon_t \sim \mathcal{N}(0,1).$$

This creates a time series with a linear upward trend. Similarly to what done before:

$$\mathbb{E}(y_t) = \mathbb{E}(\epsilon_t) + \mathbb{E}(0.1 \cdot t) = 0.1 \cdot t.$$

Again, saying that we wish to predict the behaviour of the time series at time 150, we know this will be centered on $\mathbb{E}(y_{150}) = 1.5$ (and with which variance?).

C. In the rightmost example, the time series was generated for the first half of the observations as in A., however after $t = 50$, a sudden shift occurs. Mathematically:

$$y_t = \begin{cases} \epsilon_t & \text{for } t \leq 50 \\ \epsilon_t + 5 & \text{for } t > 50 \end{cases}, \quad \epsilon_t \sim \mathcal{N}(0,1)$$

This abrupt change at $t = 50$ introduces a piecewise structure to the data, where the data is seen following a distribution prior to the change, $y_t \sim N(0, 1)$ up to a certain time point $t = 50$, and $y_t \sim N(5, 1)$ after. What can we say about time $t = 150$? Well, to make assumptions we have to assume that other changes are happening. However, it is far more interesting to find out where this change happened.
In many examples of this module, we will be studying processes that are piecewise stationary in the mean and variance, as in this example.

Stationarity in the mean and variance. A time series is said to be *stationary* in mean and variance, if its mean and variance are constant over time. That is, for a time series $y_{1:n}$:

$$\mathbb{E}(y_t) = \mu \quad \text{and} \quad \text{Var}(y_t) = \sigma^2 \quad \forall \in \{1, ..., n\}$$

Similarly, a time series is *non-stationary* in the mean and variance if those change over time.

Piecewise stationary in the mean and variance. A *piecewise stationary* time series is a special case of a non-stationary time series. We will say that a time series is *piecewise stationary* in mean and variance if it is stationary within certain segments but has changes in the mean or variance at certain points, known as *changepoints*. After each changepoint, the series may have a different mean, variance, or both.

Back to our example.

- In A., we can see, very simply how, in this case

  $$\mathbb{E}(y_t) = \mathbb{E}(\epsilon_t) = 0, \forall t \in \{1, ..., 100\},$$

  therefore our series is stationary in the mean and variance.

- In B, we notice that:

  $$\forall t_1, t_2 \in \{1, ..., 100\}, t_1 \neq t_2 \rightarrow \mathbb{E}(y_{t_1}) \neq \mathbb{E}(y_{t_2}).$$

  We can therefore say that the series is non-stationary in the mean.

- In C, $E[y_t] = E[\epsilon_t] = 0$ for $t \leq 50$, and $E[y_t] = E[\epsilon_t] + E[3] = 5$ for $t > 50$. The series is therefore piecewise stationary in the mean.

1.2 Introduction to changepoints

Changepoints are sudden, and often unexpected, shifts in the behavior of a process. They are also known as breakpoints, structural breaks, or regime switches. The detection of changepoints is crucial in understanding and responding to changes in various types of time series data.

The primary objectives in detecting changepoints include:

- **Has a change occurred?**: Identifying if there is a shift in the data.
- **If yes, where is the change?**: Locating the precise point where the change happened.
- **What is the difference between the pre and post-change data?** This may reveal the type of change, and it could indicate differences in parameter values before and after the change.
- **How certain are we of the changepoint location?**: Assessing the confidence in the detected changepoint.
- **How many changes have occurred?**: Identifying multiple changepoints and analyzing each one for similar characteristics.

Changepoints can be found in a wide range of time series, not limited to physical, biological, industrial, or financial processes, and which objectives to follow depends on the type of the analysis we are carrying.

In changepoint detection, there are two main approaches: **online** and **offline** analysis. In applications that require **online analysis**, the data is processed as it arrives, or in small batches. The primary goal of online changepoint detection is to identify changes as quickly as possible, making it crucial in contexts such as process control or intrusion detection, where immediate action is necessary.

On the other hand, **offline analysis** processes all the data at once, typically after it has been fully collected. The aim here is to provide an accurate detection of changepoints, rather than a rapid one. This approach is common in fields like genome analysis or audiology, where the focus is on understanding the structure of the data post-collection.

To give few examples:

1. **Spectroscopy data.** Changepoint detection is useful in spectroscopy data to segment time series of electron emissions into regions of approximately constant intensity, accounting for large-scale fluctuations in laser power and beam pointing.

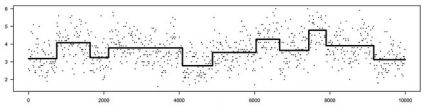


Figure 1.1: Electron emission spectroscopy data, Frick, K., Munk, A., & Sieling, H. (2014).

2. **ECG**: Detecting changes or abnormalities in electrocardiogram (ECG) data can help in diagnosing heart conditions.

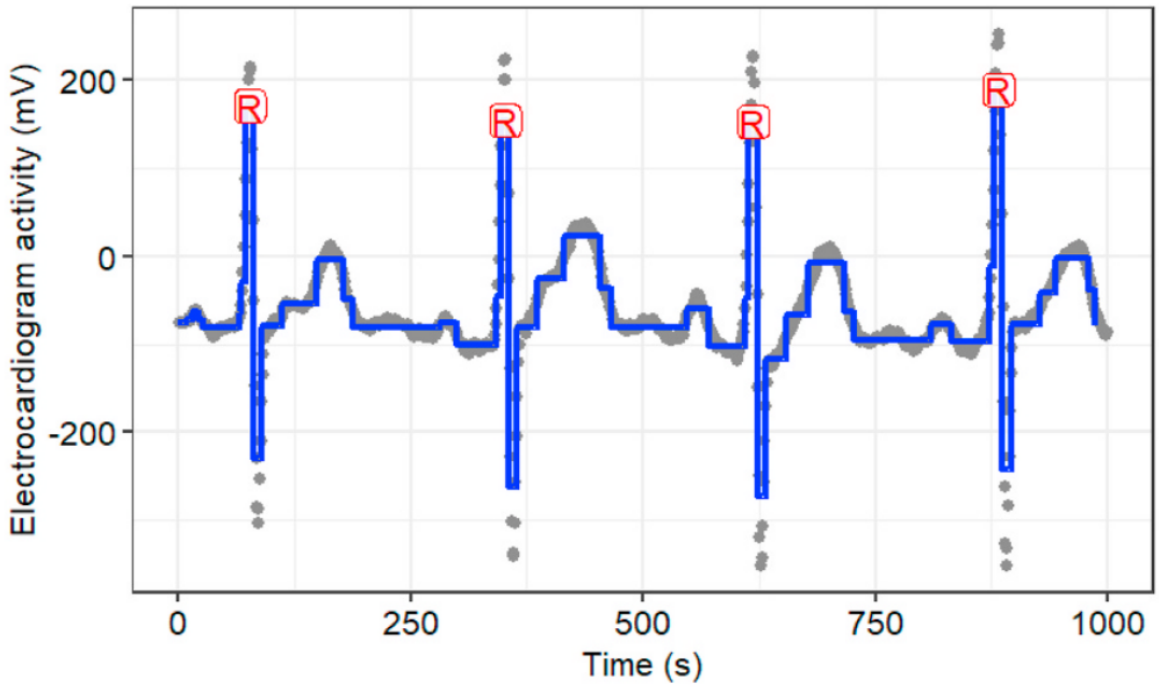


Figure 1.2: Electrocardiograms (heart monitoring), Fotoohinasab et al, Asilomar conference 2020.

3. **Cancer Diagnosis**: Identifying breakpoints in DNA copy number data is important for diagnosing some types of cancer, such as neuroblastoma. This is a typical example of an offline analysis.

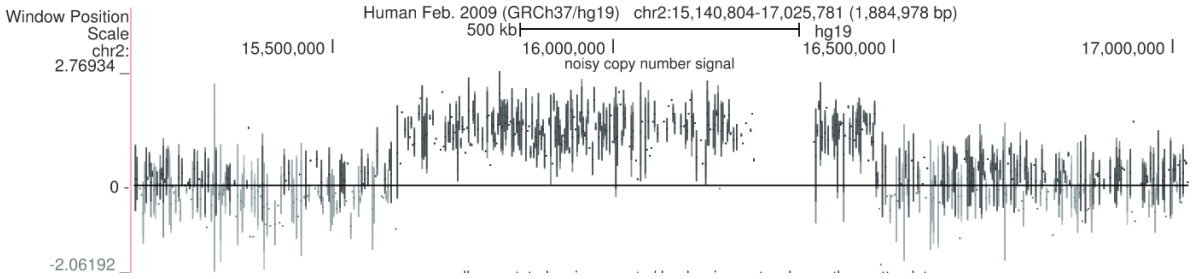


Figure 1.3: DNA copy number data, breakpoints associated with aggressive cancer, Hocking et al, Bioinformatics 2014.

4. **Engineering Monitoring**: Detecting changes in CPU monitoring data in servers can help in identifying potential issues or failures: this is often analysed in real-time on with online methods, with the aim of detecting an issue as quickly as possible.

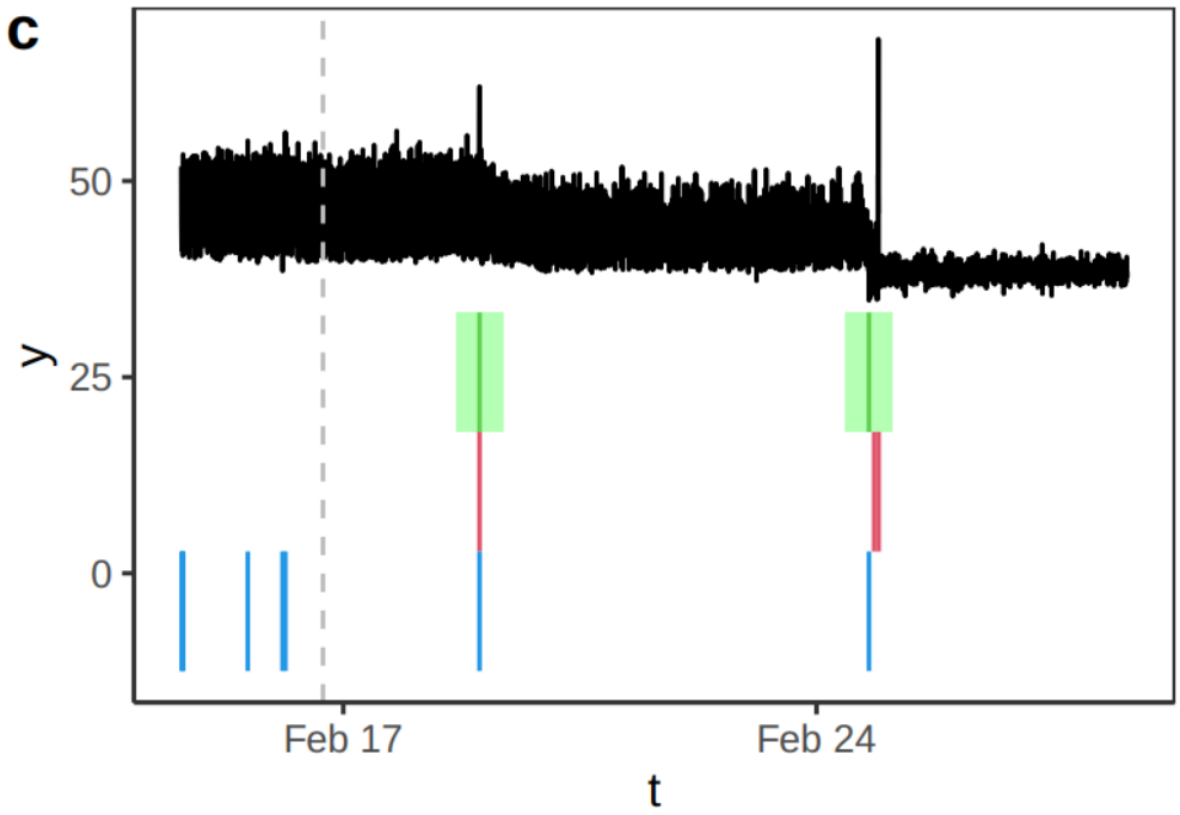


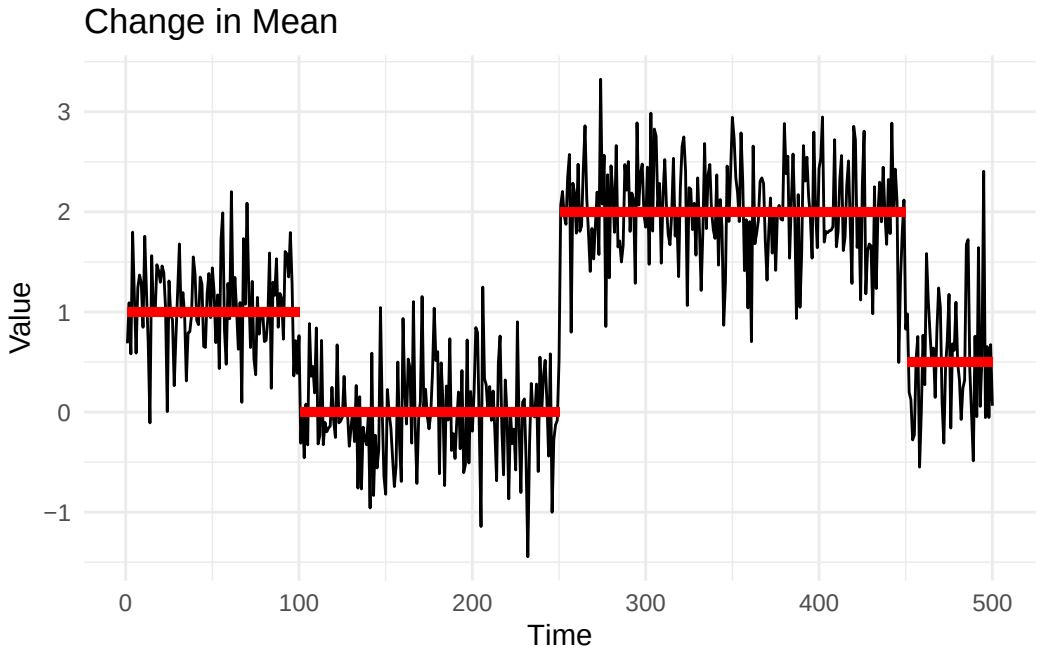
Figure 1.4: Temperature data from a CPU of an AWS server. Source Romano et al., (2023)

In this module, we will focus exclusively on **offline** changepoint detection, where we assume that all the data is available for analysis from the start.

1.2.1 Types of Changes in Time Series

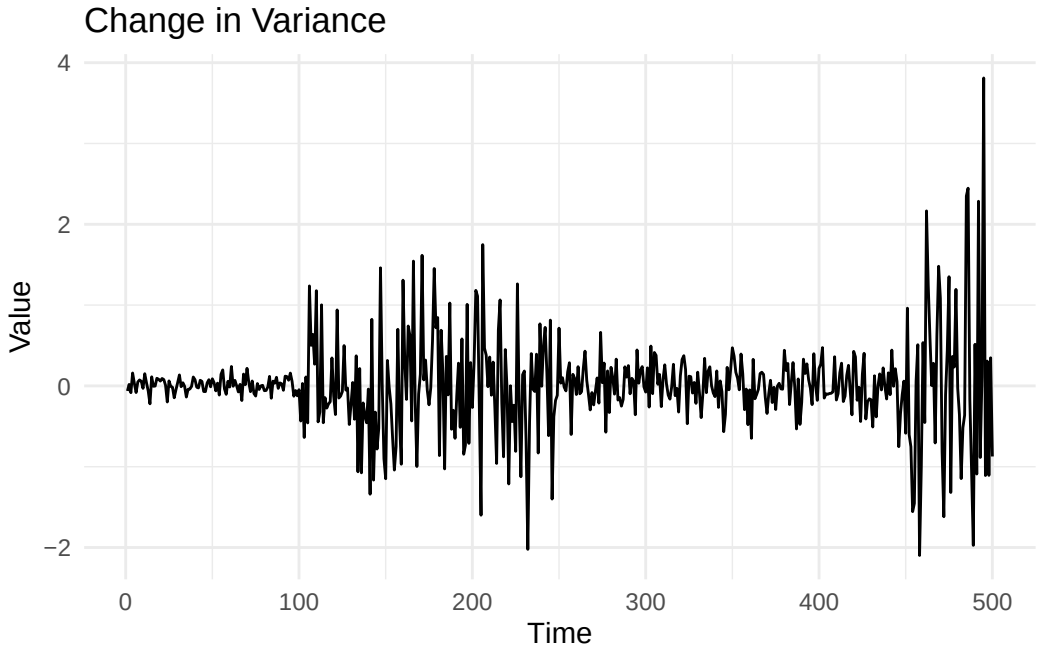
Depending on the model, we could seek for different types of changes in the structure of a time series. Some of the most common types of changes include shifts in mean, variance, and trends in regression. For example, the CPU example above exihibited, in addition to some extreme observations, both changes in mean and variance.

- A **change in mean** occurs when the average level of an otherwise stationary time series shifts from one point to another. This type of change is often encountered in real-world data when there is a sudden shift in the process generating the data, such as a change in policy, market conditions, or external factors affecting the system.



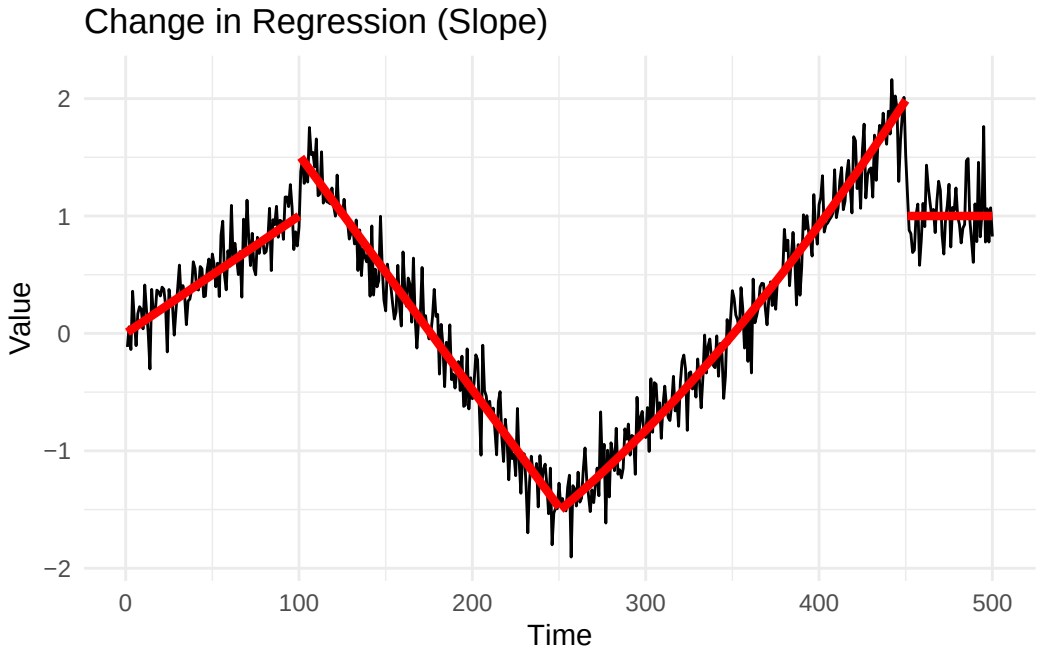
In the plot above, the red lines indicate the true mean values of the different segments.

- A **change in variance** refers to a shift in the variability of the time series data, even when the mean remains constant. This type of change is important in scenarios where the stability of a process fluctuates over time. For example, in financial markets, periods of high volatility (high variance) may be followed by periods of relative calm (low variance).



1.2.1.1 3. Change in Regression (Slope)

A **change in regression** or slope occurs when the underlying relationship between time and the values of the time series changes. This could reflect a shift in the growth or decline rate of a process. For example, a company's revenue might grow steadily over a period, then plateau, and later exhibit a quadratic or nonlinear growth trend.

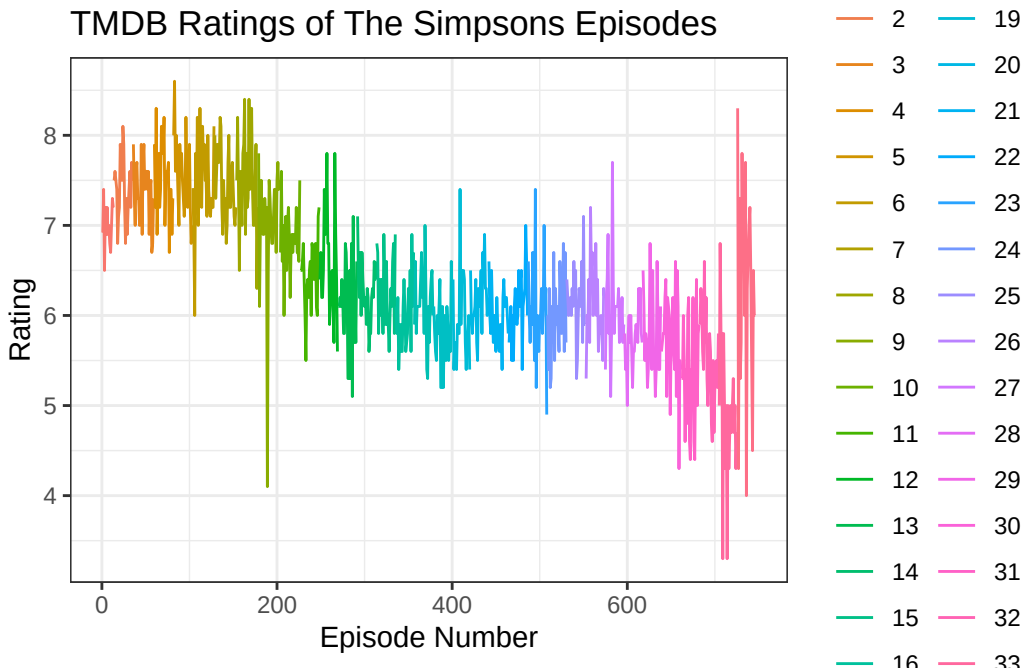


1.2.2 The biggest data challenge in changepoint detection

One of the most widely debated and difficult data challenges in changepoint detection may not be in the field of finance, genetics, or climate science—but rather in television history. Specifically, the question that has plagued critics and fans alike for years is: **At which episode did "The Simpsons" start to decline?**

It's almost common knowledge that "The Simpsons," the longest-running and most beloved animated sitcom, experienced a significant drop in quality over time. But pinpointing exactly *when* this drop occurred is the real challenge. Fortunately, there's a branch of statistics that was practically built to answer questions like these!

I have downloaded a dataset (Bown 2023) containing ratings for every episode of "The Simpsons" up to season 34. We will analyze this data to determine if and when a significant shift occurred in the ratings, which might reflect the decline in quality that so many have observed.



In this plot, each episode of "The Simpsons" is represented by its TMBD rating, and episodes are colored by season. By visually inspecting the graph, we may already start to see some potential points where the ratings decline. However, the goal of our changepoint analysis is to move beyond visual inspection and rigorously detect the exact moment where a significant shift in the data occurs.

Jokes apart, this is a challenging time series! First of all, there's not a clear single change, but rather an increase, followed by a decline. After which, the sequence seems rather stationary. For this reason, throughout the module, we will use this data as a running example to develop our understanding of various methods, hopefully trying to obtain a definitive answer towards the final chapters. But let's proceed with order…

1.3 Detecting one change in mean

In this section, we will start by exploring the simplest case of a changepoint detection problem: **detecting a change in the mean** of a time series. We assume that the data is generated according to the following model:

$$y_t = \mu_t + \epsilon_t, \quad t = 1, \dots, n,$$

where $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$ represents Gaussian noise with mean 0 and known variance σ^2, and $\mu_t \in \mathbb{R}$ is the signal at time t. The vector of noise terms $\epsilon_{1:n}$ is often referred to as Gaussian noise,

and hence, this model is known as *the signal plus noise model*, where the signal is given by $\mu_{1:n}$ and the noise by $\epsilon_{1:n}$.

In *the change-in-mean problem*, our goal is to determine whether the signal remains constant throughout the entire sequence, or if there exists a point τ, where the mean shifts. In other words, we are testing whether

$$\mu_1 = \mu_2 = \cdots = \mu_n \quad \text{(no changepoint)},$$

or if there exists a time τ such that

$$\mu_1 = \mu_2 = \cdots = \mu_\tau \neq \mu_{\tau+1} = \cdots = \mu_n \quad \text{(changepoint at } \tau\text{)}.$$

Note. The point τ is our *changepoint*, e.g. the first point after which our mean changes, however there's a lot of inconsistencies on the literature: sometimes you will find that people refer to $\tau+1$ as the changepoint, and τ as the last pre-change point (as a matter of fact, please let me know if you spot this inconsistency anywhere in these notes!).

To address this problem, one of the most widely used methods is *the CUSUM (Cumulative Sum) statistic*. The basic idea behind the CUSUM statistic is to systematically compare the mean of the data to the left and right of each possible changepoint τ. By doing so, we can assess whether there is evidence of a significant change in the mean at a given point.

1.3.1 The CUSUM statistics

The CUSUM statistic compares, for a fixed $\tau \in \{1, \dots, n-1\}$, the empirical mean (average) of the data to the left (before τ) with the empirical mean of the data to the right (after τ):

$$C_\tau = \sqrt{\frac{\tau(n-\tau)}{n}} \left| \bar{y}_{1:\tau} - \bar{y}_{(\tau+1):n} \right|,$$

Our $\bar{y}_{1:\tau}$ and $\bar{y}_{(\tau+1):n}$ are just the empirical means of each segment, simply computed with:

$$\bar{y}_{l:u} = \frac{1}{u-l+1} \sum_{t=l}^{u} y_t.$$

The term on the left of the difference, is there to re-scale it so that our statistics is the absolute value of normal random variable that has variance σ^2. If there is no change at τ, this difference is going to be distributed as a standard normal.

This approach is intuitive because if the mean μ is the same across the entire sequence, the values of the averages on both sides of any point τ should be similar. However, if there

is a large-enough change in the mean, the means will differ significantly, highlighting the changepoint.

More formally, we declare a change at τ if:

$$\frac{C_\tau^2}{\sigma^2} > c,$$

where the $c \in \mathbb{R}^+$ is a suitable chosen threshold value (in fact it is often chosen as in hypothesis testing).

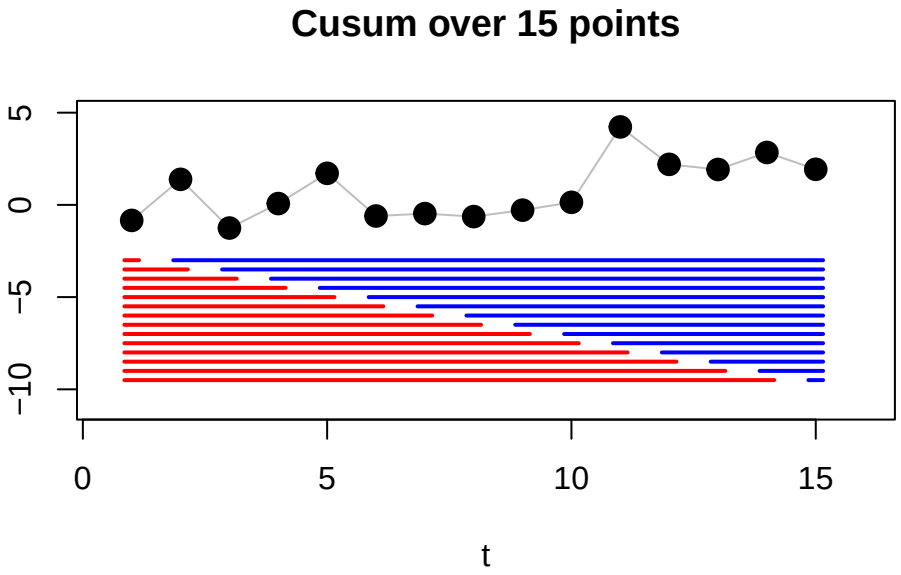
1.3.2 Searching for all τs

In practice, however, we do not know the changepoint location in advance. Our goal is to detect whether a changepoint exists and, if so, estimate its location. To achieve this, we need to consider all possible changepoint locations and choose the one that maximizes our test statistic.

The natural extension of the CUSUM to this situation is to use as a test statistic the maximum of C_τ as we vary τ:

$$C_{max}^2 = \max_{\tau \in \{1,\dots,n-1\}} C_\tau^2/\sigma^2.$$

And detect a changepoint if $C_{max}^2 > c$ for some suitably chosen threshold c. The choice of c will determine the significance level of the test (we'll discuss this in more detail later). Graphically, the test will look as follows:



If we detect a changepoint (i.e., if $C^2_{max} > c$), we can estimate its location by:

$$\hat{\tau} = \arg \max_{\tau \in \{1,\dots,n-1\}} C^2_\tau.$$

In other words, $\hat{\tau}$ is the value of τ that maximizes the CUSUM statistic.

A simple estimate of the size of the change is then given by:

$$\Delta\hat{\mu} = \bar{y}_{(\hat{\tau}+1):n} - \bar{y}_{1:\hat{\tau}}.$$

This estimate represents the difference between the mean of the data after the estimated changepoint and the mean of the data before the estimated changepoint.

1.3.3 Example

Let us compute the cusum for the vector $y_{1:4} = (0.5, -0.1, 12.1, 12.4)$.

We know that $n = 4$ (the total number of observations), therefore possible changepoints are: $\tau = 1, 2, 3$.

Compute empirical means for each segment

We first need to calculate the segment means, $\bar{y}_{1:\tau}$ and $\bar{y}_{(\tau+1):n}$, for each τ.

- For $\tau = 1$, the left segment is: $y_{1:1} = (0.5)$, and $\bar{y}_{1:1} = 0.5$. The right segment: $y_{2:4} = (-0.1, 12.1, 12.4)$ gives $\bar{y}_{2:4} = \frac{-0.1+12.1+12.4}{3} = \frac{24.4}{3} = 8.13$.
- For $\tau = 2$, we have, in a similar fashion, $\bar{y}_{1:2} = \frac{0.5-0.1}{2} = 0.2$, $\bar{y}_{3:4} = \frac{12.1+12.4}{2} = 12.25$,
- Lastly, for $\tau = 3$, we have $\bar{y}_{1:3} = \frac{0.5-0.1+12.1}{3} = \frac{12.5}{3} = 4.16$ and $\bar{y}_{4:4} = 12.4$.

Compute the CUSUM statistics

Now that we have the empirical means for each segment, we have all the ingredients for computing our CUSUM:

$$C_\tau = \sqrt{\frac{\tau(n-\tau)}{n}} \left| \bar{y}_{1:\tau} - \bar{y}_{(\tau+1):n} \right|.$$

- **For** $\tau = 1$:
$$C_1 = \sqrt{\frac{1(4-1)}{4}} \left| 0.5 - 8.13\overline{3} \right| = 0.866 \times 7.63\overline{3} = 6.61.$$
- **For** $\tau = 2$:
$$C_2 = \sqrt{\frac{2(4-2)}{4}} \left| 0.2 - 12.25 \right| = 1 \times 12.05 = 12.05.$$

- **For** $\tau = 3$:

$$C_3 = \sqrt{\frac{3(4-3)}{4}}\left|4.16\overline{6} - 12.4\right| = 0.866 \times 8.23\overline{3} = 7.13.$$

Thus, the maximum of the CUSUM statistic occurs at $\tau = 2$, with $C_{max} = 12.05$. To detect a changepoint, we would compare C_{max} to a threshold value c. If $C_{max} > c$, we conclude that there is a changepoint at $\hat{\tau} = 2$.

1.3.4 Algorithmic Formulation of the CUSUM Statistic

This process seems rather long, as for every step, we need to precompute the means… A naive implementation of the cusum, in fact, takes $\mathcal{O}(n^2)$ computations.

However, there's an algorithmic trick: by sequentially computing partial sums, e.g. $S_n = \sum_{i=1}^{n} y_i$, we can shorten out our computations significantly. In this way we can compute the value of the means directly as we iterate in the for cycle.

INPUT: Time series $y = (y_1, ..., y_n)$, threshold c, variance σ.
OUTPUT: Changepoint estimate $\hat{\tau}$, maximum CUSUM statistic C_{max}

$n \leftarrow$ length of y
$C_{max} \leftarrow 0$
$\hat{\tau} \leftarrow 0$
$S_n \leftarrow \sum_{i=1}^{n} y_i$ // Compute total sum of y
$S \leftarrow 0$

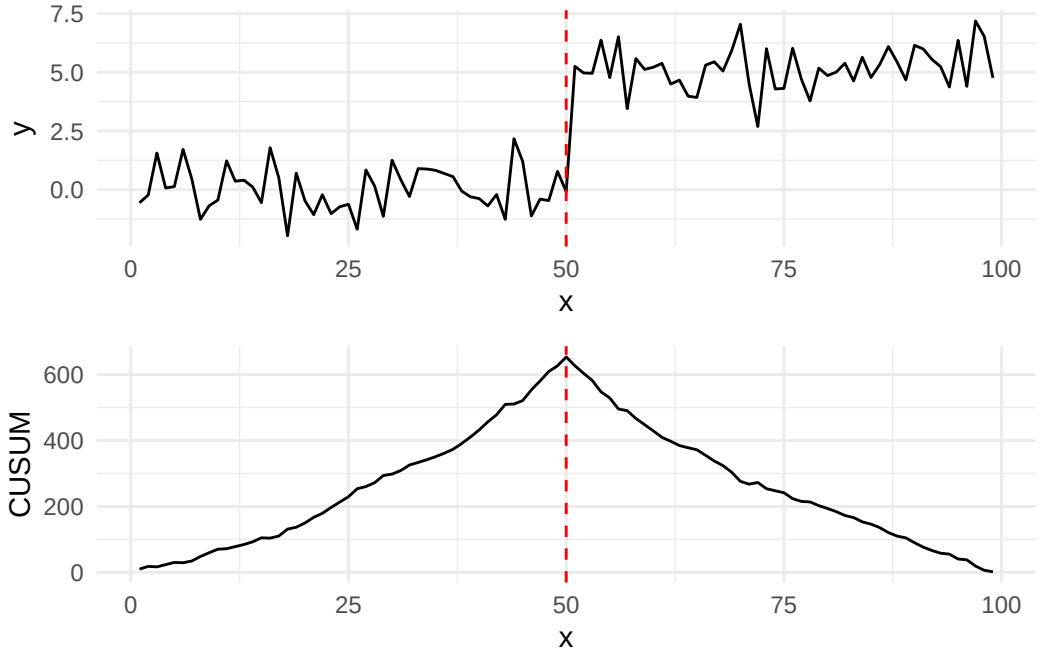
FOR $t = 1, \dots, n-1$
  $S \leftarrow S + y_t$
  $\bar{y}_{1:t} \leftarrow S/t$
  $\bar{y}_{(t+1):n} \leftarrow (S_n - S)/(n-t)$ // Can you figure out why?
  $C_t^2 \leftarrow \frac{t(n-t)}{n}(\bar{y}_{1:t} - \bar{y}_{(t+1):n})^2$
  IF $C_t^2 > C_{max}$
    $C_{max} \leftarrow C_t$
    $\hat{\tau} \leftarrow t$

IF $C_{max}/\sigma^2 > c$
  RETURN $\hat{\tau}$, C_{max} // Changepoint detected
ELSE
  RETURN NULL, C_{max} // No changepoint detected

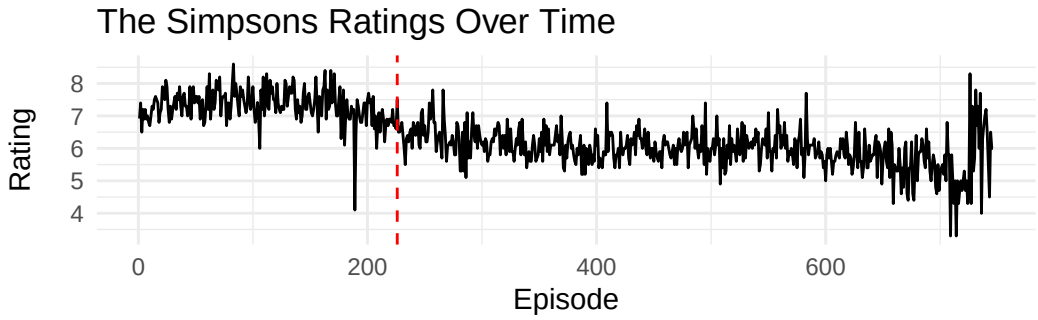
For this reason, the time complexity of the CUSUM algorithm is $O(n)$, where n is the length of the time series.

1.3.5 Example: a large sequence

We can see how the cusum behaves across different values of $\tau = 1, \ldots, n-1$ in the example below:



Running the CUSUM test, and maximising on our Simpsons episode, results in:



This results in episode Thirty Minutes over Tokyo being the last "good" Simpsons episode, with Beyond Blunderdome being the start of the decline, according to the Gaussian change-in-mean model!

1.4 Exercises

1.4.1 Code the CUSUM algorithm for a unknown change location, based on the pseudocode above.

Workshop 1

1. Determine if the following processes are stationary, piecewise stationary, or non-stationary:

a. $y_t = y_{t-1} + \epsilon_t, \quad t = 2, \dots, n, y_1 = 0, \epsilon_t \sim N(0,1)$. This is a random walk model. Let's start by computing the expected value and variance of y_t across all t. **TIP:** Start by expanding y_t in terms of the noise components…

b. $y_t = t\epsilon_t + 3\mathbb{1}(t > 50), \quad t = 1, \dots, 100, \quad \epsilon_t \sim N(0,1)$

c. $y_t = 0.05 \cdot t + \epsilon_t, \ t = 1, \dots, 100, \quad \epsilon_t \sim N(0,1)$

2. In this exercise we will show that:

$$\frac{1}{\sigma^2}\sqrt{\frac{\tau(n-\tau)}{n}}(\bar{y}_{1:\tau} - \bar{y}_{(\tau+1):n})$$

follows a standard normal distribution. **Hint:**

   a. Compute the expected value and variance of the difference $\bar{y}_{1:\tau} - \bar{y}_{(\tau+1):n}$

   b. Conclude that if you standardise the sum, this follows a standard normal distribution.

1.4.2 Lab 1

1. Code the CUSUM algorithm for a unknown change location, based on the pseudocode of Section Section **??**.
2. Modify your function above to output the CUSUM statistics over all ranges of tau.
3. Recreate the "CUSUM Statistics over time" plot for the Simpsons data above.

   a. You'll be able to load the dataset via:

```
library(tidyverse)
simpsons_episodes <- read_csv("https://www.lancaster.ac.uk/~romano/teaching/2425MATH337/datas

simpsons_ratings <- simpsons_episodes |>
  mutate(Episode = id + 1, Season = as.factor(season), Rating = tmdb_rating)
simpsons_ratings <- simpsons_ratings[-nrow(simpsons_ratings), ]

# run your CUSUM algorithm on the Rating variable!
```

   b. To run it on the whole sequence, you'll have to set the threshold $c = \infty$.

   c. Assume $\sigma^2 = 1$

2 Controlling the CUSUM and Other Models

In this chapter, we explore the properties of the CUSUM test for detecting a change in mean, and this will allow us how to determine appropriate thresholds, and explore its properties when a changepoint is present.

We will employ some concepts from asymptotic theory: in time series analysis, an asymptotic distribution refers to the distribution that our test statistic approaches as the length of the time series n becomes very large.

2.1 The asymptotic distribution of the CUSUM statistics

If $z_1, \cdots, z_k$ are independent, standard Normal random variables, then:

$$\sum_{i=1}^{k} z_i^2 \sim \chi_k^2,$$

where χ_k^2 is a chi-squared distribution with k degrees of freedom. The chi-squared distribution is a continuous probability distribution that models the sum of squares of k independent standard normal random variables: we have met the chi-squared distribution already in hypothesis testing and constructing confidence intervals. The shape of the distribution depends on its degrees of freedom. For $k = 1$, it's highly skewed, but as k increases, it becomes more symmetric and approaches a normal distribution.

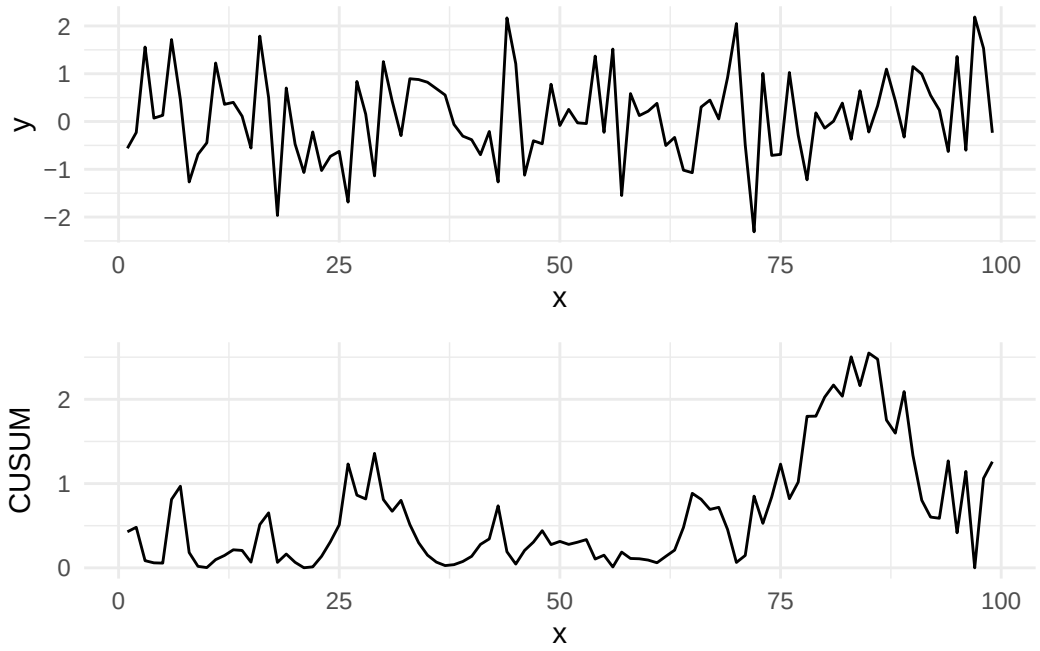
Last week, we found out that, under the null hypothesis of no change:

$$\frac{1}{\sigma^2}\sqrt{\frac{\tau(n-\tau)}{n}}(\bar{y}_{1:\tau} - \bar{y}_{(\tau+1):n}) \sim N(0,1).$$

Therefore, our test statistics for a fixed τ:

$$\frac{C_\tau^2}{\sigma} \sim \chi_1^2.$$

If we take the example of last week, and remove the changepoint, we can observe that the cusum statistics stays constant, and relatively small:



However, as the change is unknown, our actual test statistic for detecting a change is $\max_\tau C_\tau^2/\sigma^2$.

For this reason, calculating the distribution of this maximum ends up being a bit more challenging…

1. So far, we only studied the behaviour of the statistics for one fixed τ, however, when comparing the maximums, the values of C_τ are in fact not independent across different τs.
2. As we will learn later, the CUSUM is a special case of a LR test, as setting the size of the actual change in mean to 0 effectively removes the changepoint parameter from the model. For this reason, the usual regularity conditions for likelihood-ratio test statistics don't apply here.

2.1.1 Controlling the max of our cusums

Fortunately, for controlling our CUSUM test, we can use the fact that $(C_1, ..., C_{n-1})/\sigma$ are the absolute values of a Gaussian process with mean 0 and known covariance, and there are well known statistical results that can help us in our problem. Yao and Davis (1986), in fact, show that the maximum of a set of Gaussian random variables is known to converge to a Gumbel distribution, described by the following equation:

$$\lim_{n\to\infty} \Pr\{a_n^{-1}(\max_\tau C_\tau/\sigma - b_n) \leq u_\alpha\} = \exp\{-2\pi^{-1/2}\exp(-u_\alpha)\}, \tag{2.1}$$

where $a_n = (2 \log\log n)^{-1/2}$ and $b_n = a_n^{-1} + 0.5 a_n \log\log\log n$ are a scaling and a centering constant.

The right side of this equation is the CDF of a Gumbell distribution. As we learned from likelihood inference, to find the threshold c_α for a given false probability rate, we first set the right-hand side equal to $1 - \alpha$, and solve for u_α. This gives:

$$u_\alpha = -\log\left(-\frac{\log(1-\alpha)}{2\pi^{-1/2}}\right).$$

Then, we can find the critical value by looking into the left side of the equation:

$$\tilde{c} = (a_n u_\alpha + b_n),$$

To find the threshold, as $\max_\tau \frac{C_\tau^2}{\sigma^2} > c$, we just have to square our value above, e.g. $c_\alpha = \tilde{c}^2$.

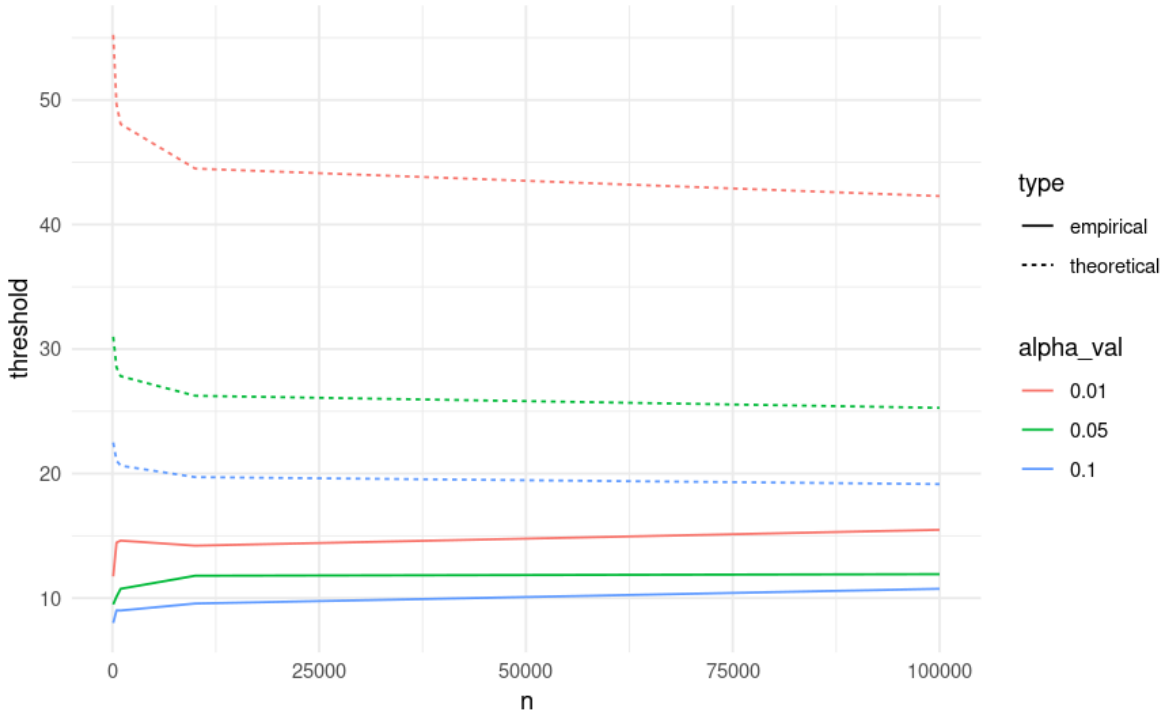
This asymptotic result suggests that the threshold c_α for C_τ^2/σ^2 should increase with n at a rate of approximately $2 \log\log n$. Given that this is a fairly slow rate of convergence, this suggests that the threshold suggested by this asymptotic distribution can be conservative in practice, potentially leading to detect less changepoints than what actually exist.

In practice, it's often simplest and most effective to use Monte Carlo methods to approximate the null distribution of the test statistic. This can be done via the following process:

1. Simulate many time series under the null hypothesis (no changepoint),
2. Calculate the test statistic C_τ^2/σ^2 for each one of the replicates.
3. Set the threshold to be the $(1 - \alpha)$ percentile of the distribution of the test statistics from simulated data.

This leads to have less conservative thresholds.

Theoretical vs Empirical Thresholds The figure below shows, for various levels of $\alpha = 0.01, 0.05, 0.1$, thresholds c_α computed from the theoretical distribution of Equation **??** against the Monte Carlo thresholds obtained from empirical simulations under the null.



We will see how to compute in practice the theoretical and empirical thresholds in the Lab!

2.2 The Likelihood Ratio Test

The CUSUM can be viewed as a special case of a more general framework based on the Likelihood Ratio Test (LRT). This allow us to test for more general settings, beyond simply detecting changes in the mean.

In general, the Likelihood Ratio Test is a method for comparing two nested models: one under the null hypothesis, which assumes no changepoint, and one under the alternative hypothesis, which assumes a changepoint exists at some unknown position τ.

Suppose we have a set of observations $x_1, x_2, \dots, x_n$. Under the null hypothesis H_0, we assume that all the data is generated by the same model without a changepoint. Under the alternative hypothesis H_1, there is a single changepoint at τ, such that the model for the data changes after τ. The LRT statistic is given by:

$$LR_\tau = -2 \log \left\{ \frac{\max_\theta \prod_{t=1}^{n} f(y_t|\theta)}{\max_{\theta_1,\theta_2} [(\prod_{t=1}^{n} f(y_t|\theta_1))(\prod_{t=1}^{n} f(y_t|\theta_2)]} \right\} \tag{2.2}$$